Introduction to data warehousing and data mining

Suyog Dhokpande, Hitesh raut

Abstract— The Data Warehousing supports business analysis and decision making by creating an enterprise wide integrated database of summarized, historical information. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.
This paper describes about the basic architecture of data warehousing, its software and process of data warehousing. It also presents different techniques followed in data mining.

Index Terms- data warehousing, data mining, clusture, OLAP, EIS.

1 INTRODUCTION

It is the repository of data that are organized by subject to support decision makers in the organizations. It is the concept was intended to provide architectural model for then flow of data from operational system to the decision support environment. A common source of data for data warehouse is nothing but the operational data base of the companies

data warehouse can be define as "a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process".

A data warehouse provides a mechanism for separating operational and informational processing, with information being the domain of the warehouse. Since the warehouse is populated by data created by the operational environment, the flow of information is usually one way ,from operation data stores to the data warehouse.

In the absence of data warehousing architecture, an enormous amount of redundancy was required to support environments. In larger corporation, it is difficult for multiple support environment to work independently. The process of gathering, cleaning and integrating data from various source, usually from long term existing operational system (legacy system) was typically in part replicated for each environment.

2 CHARECTERISTIC OF DATA WAREHOUSING

2.1 **Subject-oriented:** data is organized according to subject instead of application

2.2 **Integrated:** When data resides in many separate applications in the operational environment, encoding of data is often inconsistent.

2.3 Time-variant: The data warehouse contains data for comparison.

2.4 **Non-volatile:** Data are not updated or changed in any way once they enter the data warehouse, but are only loaded and accessed.

3 WHAT IS DATA WAREHOUSE SOFTWARE ?

Data warehouse software has grown exponentially in the past several years and is expected to experience above average growth well into the future. A data warehouse is a repository of all the transactional data of an organization or company. The primary purpose of a data warehouse is to analyze transactions and run complex reports.

There are three primary functions to every data warehouse software product:

1. data extracting,
2. creating the database structure,
3. creating customized queries.

In the information technology industry, data warehouse experts are known as business intelligence specialists. They typically have a background in math, statistics, or computer system analysis. Additional training is often required in relational databases, system architecture, and the fundamentals of database programming.

One of the most important functions of any data warehouse software is the ability to correctly extract and structure data from a variety of sources. This is often called an extract, transform, and load (ETL) tool. Data warehouses must be populated with data from the transaction system in a way that maintains the integrity and inter-relationships of the data, while allowing the staff to customize the data being extracted. This is an essential part of the architecture of the system.

It is important to note that most data warehouse software programs are used to create, support, and maintain multiple data sets.

4 DATA WAREHOUSE ARCHITECTURE

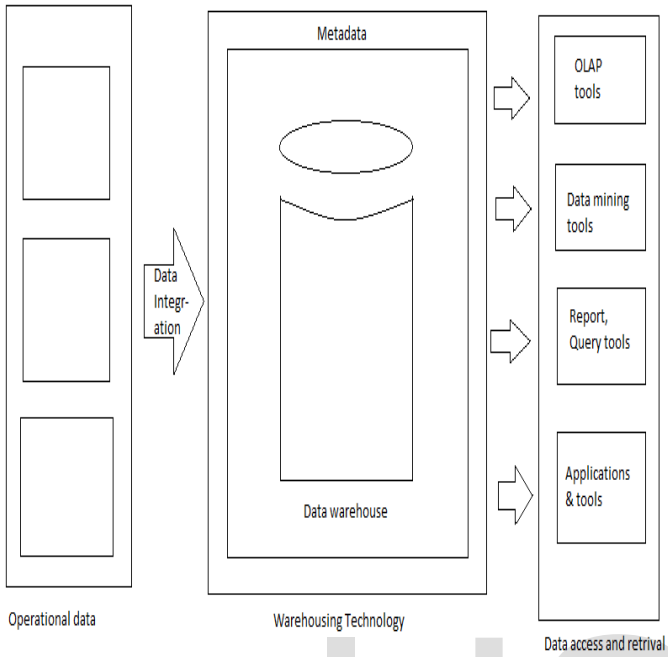

Fig. Data warehouse Architecture

4.1 Operational data: The source data for warehouse is the operational applications. In most organizations you will find really large database in operation for normal daily transactions. These types of databases are known as Operational database.

4.2 Data Integration: Extracting data from source system, transfer them, cleaning and load them into data marts or data warehouse is called data integration.

4.3 Metadata: It is the data about data and contains the location and description of warehouse system components; names, definition, structure and contents of data warehouse and end-user views.

4.4 Data warehouse: A data warehouse is subject oriented , integrated time variant, non volatile collection of data in support of management decision. The data in data warehouse contains large historical components (covering 5 to 10 years). The data warehouse must be capable of holding and managing large volumes of data as well as different data structures for the same database over time.

4.5 Data access and data retrieval: The five main groups for data access and data retrieval are

1) Data query and reporting tools
2) Application development tools
3) Executive information System(EIS)
4) Online analytical processing tools(OLAP)
5)Data mining tools

5 DATA ARCHITECT

The data architect is responsible for identifying the shared data, confirming that it is correct, and allowing this information to be available in multiple cubes, without impacting the integrity of each cube. In addition, he or she is responsible for managing the data upload process required to refresh the data cubes. This can be an automated or manual process, depending on the complexity of the cubes and how the data is provided.

All data warehouse software programs come with a range of standard reports and queries. These reports are based on common business needs and tend to be quite general in nature. For example, a report of the top ten clients by sales volume for the current year is a common report request and would be standard in most programs.

6 STORAGE ARCHITECT

A storage architect performs an important function in the operation of an organization by building central database systems that house crucial information pertaining to such areas as compliance, finance, accounting, human resources, legal and other key areas. Storage architects use a variety of software tools and work with a diverse range of hardware components to accomplish those objectives. Included as routine duties, he or she may analyze data, determine key requirement to store and access that data, and communicate that information to either clients or management.

Thereafter, using cost-effective strategies, the engineer will plan and design storage systems, both for temporary use and to meet long-term data storage needs.

Tasks carried out by a storage architect are consistent among advertised positions; what varies is the level of responsibility. Generally, a storage architect will document design and installation specifications related to data storage systems, determine costs associated with a project, and evaluate technology solutions. He or she will may also liaison with other information technology professionals to design optimal storage solutions, identify storage needs, and guide the design and installation process. Upon completion of installation, responsibilities will also include testing the system, evaluating its usage, and monitoring the system to mitigate potential problems. When necessary, he or she may also specify and oversee changes to further optimize the solution based on usage patterns and requirements.

7 DATA MINING:

Data Mining or knowledge discovery in databases is the nontrivial extraction of implicit and previously unknown and potentially useful information from the data. Data mining is the search for relationship and global patterns that exist in large databases but are hidden among vast amount of data.

8 WORKING PROCEDURE :

Data Mining software analyzes relationships and patterns in stored transactions data based on open-ended user queries. Generally sought four types of relationships are :

1 **classes :** Stored data is used to locate data in predetermined groups.
2 Clusters : Data items are grouped according to logical relationships or consumer preferences.
3 Associations : Data can be mined to identify associations.
4 Sequential patterns : Data is mined to anticipate behavior patterns and trends.

Major Steps :

Extract, transform and load transaction data onto the data warehouse system. Store and manage the data in a multidimensional database system. Provide data access to business analysts and Information technology professionals. Analyze the data by application software. Present data in useful manner such as graph or table.

9 DATA MINING TECHNIQUES:

These provide a description of some of the most common data mining algorithms in use today. We have broken the discussion into two sections, each with a specific theme:

Classical Techniques: Statistics, Neighborhoods and Clustering

Next Generation Techniques: Trees, Networks and Rules

9.1 Classical Techniques :

1.**Statistics :** By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications.

2.Nearest Neighbor : Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. Nearest neighbor is a prediction technique that is quite similar to clustering -

3.Clustering : Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation -

TABLE 1
COMMERCIALLY AVAILABLE CLUSTER TAGS

Name	Income	Age	Education	Vendor
Blue Blood Estates	High	35-54	College	Claritas PRIZM*
Shotguns and Pickups	Middle	35-64	High School	Claritas PRIZM*
Southside City	Low	Mix	Grade School	Claritas PRIZM*
Living of the land	Middle-low	Families with School-age children	Low	Equifax MicroVision*
University USA	Very low	Young mix	Medium high	Equifax MicroVision*

Table Some Commercially Available Cluster Tags
This clustering information is then used by the end user to tag the customers in their database. Once this is done the business user can get a quick high level view of what is happening within the cluster. Once the business user has worked with these codes for some time they also begin to build intuitions about how these different customers clusters will react to the marketing offers particular to their business.

9.2 Next Generation Techniques : The data mining techniques in this section represent the most often used techniques that have been developed over the last two decades of research. These techniques can be

used for either discovering new information within large databases or for building predictive models.

10 DECISION TREES :

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. For instance if we were going to classify customers who churn (don't renew their phone contracts) in the Cellular Telephone Industry a decision tree might look something like that found in Figure.

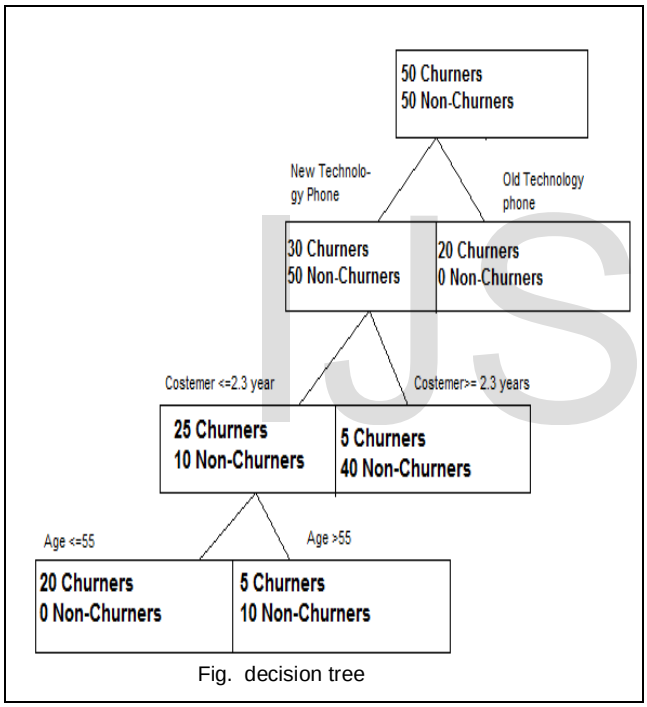

Fig. decision tree

Figure decision tree is a predictive model that makes a prediction on the basis of a series of decision much like the game of 20 questions.

We may notice some interesting things about the tree:

- It divides up the data on each branch point without losing any of the data (the number of total records in a given parent node is equal to the sum of the records contained in its two children).
- The number of churners and non-churners is conserved as you move up or down the tree

11 WHICH TECHNIQUE AND WHEN?

Some of the criteria that are important in determining the technique to be used are determined by trial and error. There are definite differences in the types of problems that are most conducive to each technique but the reality of real world data and the dynamic way in which markets, customers and hence the data that represents them is formed means that the data is constantly changing. These dynamics mean that it no longer makes sense to build the "perfect" model on the historical data since whatever was known in the past cannot adequately predict the future because the future is so unlike what has gone before.

12 POTENTIAL APPLICATIONS

Data mining has many and varied fields of application some of which are listed below.

1.Retail/Marketing
2.Banking
3.Insurance and Health Care
4.Transportation
5.Medicine

13 CONCLUSION :

Data Mining is not a new phenomenon. All large organizations already have data warehouses, but they are just not managing them. The Data Warehousing solution should enhance intelligence in decision-making process of an enterprise. Over the next few years, the growth of data mining is going to be enormous with new products and technologies coming out frequently.

In order to get the most out of this period, it is going to be important that data warehousing and mining planners and developers have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility for tomorrow.

14 REFERENCES

[1] PAPER ON DATA WAREHOUSING AND DATA MIN ING by K. Narendra (III C.S.E) V. Srinivasrao (IIIrd ECE), Chirala Engg College.

[2] PAPER ON DATA WAREHOUSING AND MINING by s.indira priya darsini ¾ C.S.E., v.chaitanya ¾ I.S.T , Koneru Lakshmaiah College Of Engineering

[3] Data warehousing, Data Mining and OLAP by Alex Berson, Stephen J. Smith (reference)

[4] Data Warehousing by Reema Thareja, IT faculty, Department of CSE, Institute of Information Technology & management, GGS IP University, New Delhi.
[5] Data Mining by Pieter Adriaans , Dolf Zantinge
[6] Decision Support and Data Warehouse Systems
by Efrem G. Mallach
[7] The SAS Data Warehouse: A Real World Example
Martin P. Bourque, SAS Institute Inc., Cary, NC
[8] www.ask.com
[9] WHAT ARE ADVANTAGES AND DISADVANTAGES OF DATA WAREHOUSES? by Dan Power Editor, DSSResources.com
[10] www.wisegeek.com
[11]www.ehow.com
[12] Data Warehouse in the Enterprise,
A Competitive Review of Enterprise Data Warehouse Appliances and Technology Solutions, SQL Server Technical Article, Published on **:** January 2009